DIGITAL COMMUNICATIONS

We have learned the basics of analog radio communications modulation, e.g., amplitude, frequency, and phase modulations (AM, FM, and PM). Modern communications systems are rapidly becoming digital in modulation. Digital and discrete are often used interchangeably to connote that the modulating signal is permitted to assume only predetermined discrete values. Having only a certain number of allowed values allows for ease of processing and, perhaps more importantly, a higher threshold of noise immunity.

Digital modulation is performed using the same basic modulation as analog, i.e., AM, FM, and PM. Having a prior understanding of these analog modulations greatly facilitates comprehending the digital since, as we will see, the digital forms are simply special case applications of the analog.

Of particular importance will be determining the bandwidth requirements to transmit and receive digital signals. Additionally, since digital signals are (by necessity) sampled, sampled signals make multiplexing a much easier task.

Another important difference between analog and digital is the characterization of receiver performance in the presence of noise. Whereas analog noise performance is entirely modeled by comparing the signal power to the noise power, using the so-called signal-to-noise ratio (SNR), digital systems are usually modeled by comparing the energy in a single bit to the energy of the noise. As would be expected, since power is proportional to energy, the two noise performance characterization methods can be related.

One last note before discussing the different digital modulation forms:. In analog, the process of imparting intelligence on the carrier is known as modulation. In digital, the process is identical in that intelligence is impressed upon the carrier (albeit discrete impressions), but this form of modulation is normally referred to as *keying*.

A. AMPLITUDE SHIFT KEYING

Perhaps the oldest form of digital modulation is amplitude shift keying (ASK), a form of amplitude modulation. It is achieved by simply turning on and off an otherwise unmodulated carrier. For this reason this modulation scheme is often called on-off keying (OOK). This modulation form is the basis for sending morse code over telegraph lines or radio links. The code was developed as a series of dots and dashes (the so-called *marks* and *spaces*).

While ASK is seldom used for modern, high speed communications systems, the terms mark and space have survived in the modern literature. Since most systems use a *binary* modulation code, the allowable symbols to represent the modulating signal are limited to 1s and 0s. This has led to the natural adaptation of the historical terms to let the term mark be synonymous with a one and space to be representative of a zero.

B. FREQUENCY SHIFT KEYING

The digital modulation counterpart to frequency modulation is frequency shift keying (FSK). The advantage of using FSK is that it is easy to generate and to demodulate. However, other than a few voice circuits, FSK finds little usage, as its noise imperviousness is not as reliable as other modulation forms.

The idea behind FSK is simple: generate one frequency for transmission to represent one state and another frequency to denote another. For binary transmission this means that one frequency is used for a mark and another for a space.

To develop a mathematical representation for a FSK signal, begin with a carrier voltage signal, given by

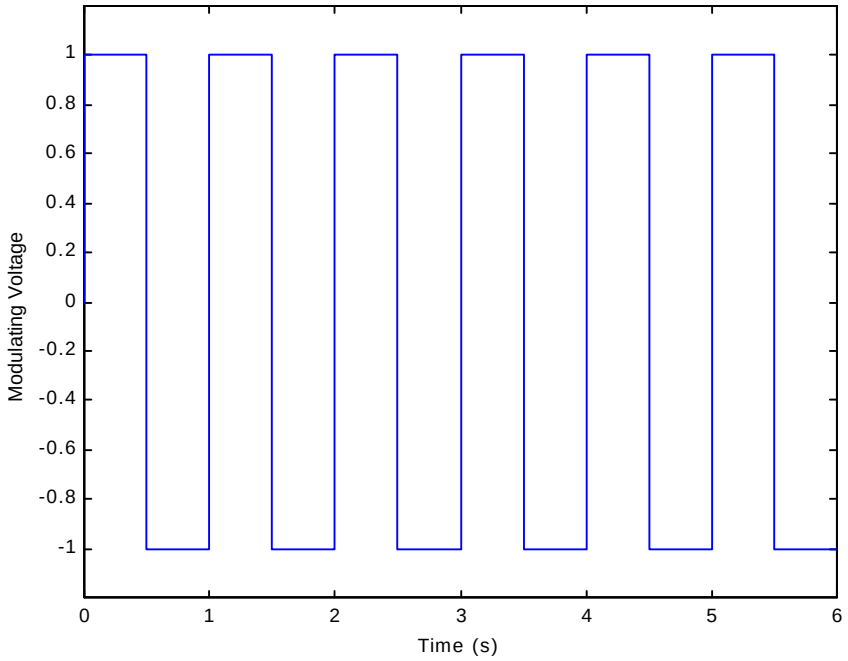
$$v(t) = V_c \cos(\omega_c t), \tag{1}$$

where v(t) is the time domain voltage signal, V_c is the peak voltage value of the carrier, and ω_c is the radian frequency of the carrier ($= 2\pi f_c$). Recall that for frequency modulation the amplitude of the modulated signal is not changed from that of the carrier, but the frequency is altered. The amount of frequency change is determined proportionally to the voltage of the modulating signal. When the modulating signal is analog, its voltage assumes an infinite number of different values, resulting in an infinite number of frequency values for the modulated FM signal.

For binary modulating signals, the voltage levels are restrained to just two values. One common voltage assignment method is to allow +1 volt to represent a one (or mark) and –1 volt to represent a zero (or space). If the modulating signal is designated as $v_m(t)$, then $v_m(t) = \pm 1$ v. Since we know that the output of the FSK modulators will be comprised of two frequencies, we can define the frequency change between the two values as δf. If the mark frequency is above the carrier frequency and the space frequency is below it, each an equal distance from the carrier, then the frequency variation can be described as $f_c \pm \delta f/2$. An appropriate equation for the FSK signal is then seen to be

$$v_{FSK}(t) = V_c \cos\left[2\pi\left(f_c + \frac{v_m(t)\delta f}{2}\right)t\right]. \quad (2)$$

This FSK signal has a constant amplitude but varies between two frequencies. The rate that the signal changes between the two frequencies is determined by the modulating signal, $v_m(t)$. For example, if $v_m(t) = \text{sign}(\sin(2\pi t))$, i.e., a square wave with period, $T_0 = 1$ second as shown in the following figure, the frequency of v(t) would change with the



switching or bit rate of the modulating signal, i.e., bit rate $= 2 f_{mod} = 2/T_0$. You can see in the figure that the polarity changes twice for every period of the input signal. If the frequency of $v_m(t)$ is changed, then the shift rate of v(t) will change as well. The rate at which the input binary signal varies is called the *bit rate*, and is given in bits per second (bps).

A pictorial representation of the output of a FSK modulator is shown in Figure 13.6 in the text. The binary input is shown as a series of ones and zeros, and directly beneath the input stream is shown the FSK output. Under the ones is shown the higher output frequency, the mark frequency f_m, while beneath the zeros is the lesser output frequency, the space frequency f_s. The top part of the figure shows the discrete frequency modulator outputs.

1. Frequency Deviation

FSK is created using a typical FM modulator. In FM we describe how the frequency deviates around the carrier frequency using the descriptor *frequency deviation*, Δf. The frequency deviation is in turn defined as

$$\Delta f = k_f A_m, \tag{3}$$

where k_f is the frequency sensitivity index and A_m is the amplitude of the modulating signal.

For FSK, k_f is constant for both mark and space inputs, and A_m has the same magnitude for each–just opposite polarity. The frequency deviation is seen to be constant in that the modulator output frequency deviates around a center carrier frequency both positive and negative by equal amounts. Therefore, it is seen that the frequency deviation is the difference between the mark and carrier frequencies or the carrier and space frequencies, i.e.,

$$\Delta f = f_m - f_c = f_c - f_s. \tag{4}$$

Using Eq. 4 to solve for the carrier frequency in terms of the mark and space frequencies, we find that the carrier is the average of the two modulator output frequencies,

$$f_c = \frac{f_m + f_s}{2}. \tag{5}$$

Inserting this value for f_c into Eq. 4 it is found that

$$\Delta f = \frac{f_m - f_s}{2}. \tag{6}$$

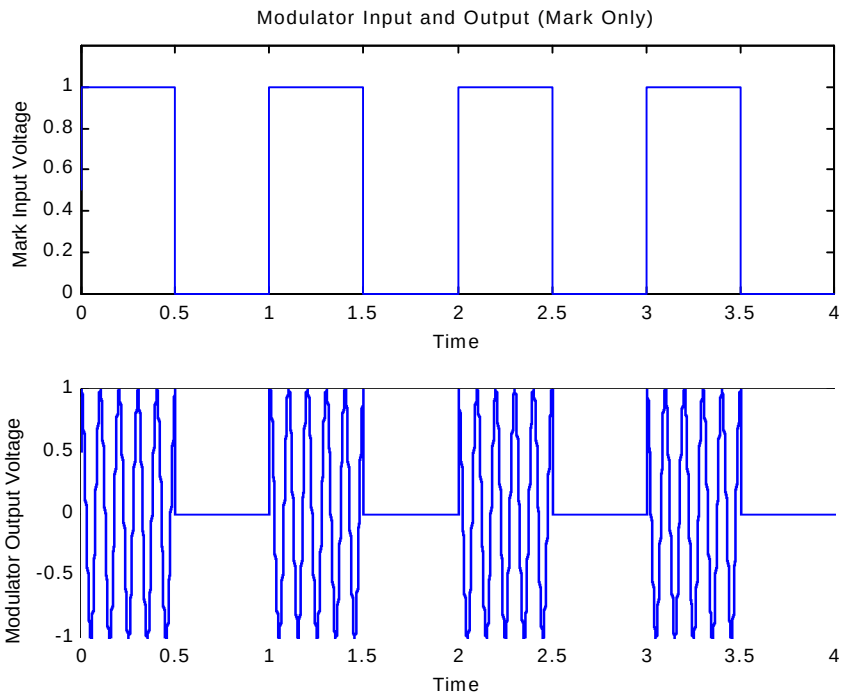
It is possible to have the mark and space frequencies reversed, i.e., the space frequency could be higher than the mark frequency. Equation 6 can be modified to include either convention as

$$\Delta f = \frac{|f_m - f_s|}{2}. \tag{7}$$

2. Bandwidth

The bandwidth required to transmit FSK is more than just the difference between the mark and space frequencies, i.e., 2Δf. Recalling the bandwidth requirements of analog FM, often estimated using Carson's rule, this increased bandwidth requirement could have been anticipated. An estimate of the bandwidth requirements for FSK can be attained using Fourier analysis.

Observe Figure 13.7, and notice that the modulator output shown in the figure can be divided into two parts: the mark frequency outputs and the space frequency outputs. Plots of the *mark* inputs and modulator outputs are shown in the figures below. As seen in the top plot, the duration of the mark bit is 0.5 seconds. The bit duration is given the name "time per bit" or t_b.



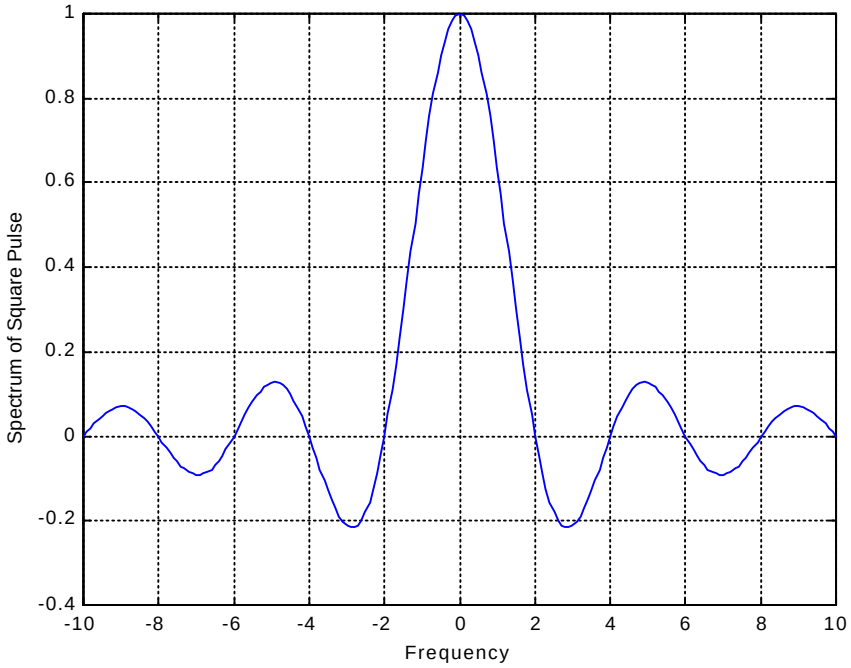
The modulator output can be seen to be the product of the binary input voltage and the mark frequency, i.e.,

$$v_{\mathrm{mod}}(t) = v_{mark}(t) \cdot \cos\left[2\pi(f_c + \Delta f)t\right] = v_{mark}(t) \cdot \cos(2\pi f_m t), \qquad (8)$$

where $v_{mod}(t)$ is the modulator output voltage and $v_{mark}(t)$ is the mark input voltage.

We wish to determine the frequency domain representation of $v_{mod}(t)$. From Fourier analysis we know that time domain multiplication of a signal $v_1(t)$ by a cosine carrier signal $v_2(t)$ results in the frequency spectrum of $v_1(t)$ being shifted out to the frequency of the carrier. Therefore the frequency domain representation of $v_{mod}(t)$ will be that of $v_{mark}(t)$ shifted out to f_m.

The problem becomes that of determining the spectrum of $v_{mark}(t)$, that is, $V_{mark}(f)$. Since $v_{mark}(t)$ is seen to be a series of square input pulses, its spectrum is a $\sin(f)/f$ waveform as shown normalized in the following figure.



This figure then represents the normalized plot of $V_{mark}(f)$.

The frequency response continues out to infinity. In order to pass all the frequency content of the signal, a system must have an infinite bandwidth. Infinite-bandwidth systems are difficult to come by so we must decide how much frequency we must allow to pass to ensure proper demodulation at the receiver. The definition of bandwidth is subjective and may be different between different systems. Recall the 3-dB bandwidth, which occurs in the above figure when the magnitude decreases to 0.707. That bandwidth definition is seldom used with this type system however. Bandwidths for square pulse signals, such as this one, are normally defined as null-to-null, i.e., from the center frequency response out to the first zero crossing. Notice where the frequency response first passes through zero. In this plot it is shown at 2 Hz. This zero crossing point can be computed in advance, it is simply $1/t_b$.

Since the frequency response of $V_{mark}(f)$ is now known, the frequency response of the system can be found easily as $V_{mark}(f - f_m)$. This action shifts the response shown in the figure above out to the mark frequency, f_m. Using the null-to-null bandwidth definition, the required bandwidth to pass the mark signal will span 4 Hz, or in general,

$$BW_{mark} = \frac{2}{t_b}. \qquad (minimum) \tag{9}$$

Now that we have determined the bandwidth requirement for the mark frequency, we can construct an identical argument for the space frequency. We will find the exact conditions for the space frequency as the mark and the bandwidth requirement will be identical to Eq. 9. The only difference is that the waveform shown in the figure above will be centered around the space frequency rather than at the mark frequency.

For proper demodulation we normally require the bandwidths of the two signals to not overlap. This necessity requires that the mark frequency be located at least $1/t_b$ above the carrier frequency and the space frequency be at least $1/t_b$ below the carrier frequency. These two requirements give

$$f_m - f_s \geq \frac{2}{t_b}. \tag{10}$$

This specifies the minimum the mark and space frequencies must be separated. The total system bandwidth extends beyond these two frequencies by $1/t_b$ on each end. The total required bandwidth is then BW = $4/t_b$ (minimum). This is pictured in Fig. 13.8 in the text.

3. Minimum Shift Keying

The transition from the space frequency to the mark frequency, and vice-versa, shown in the right side of Fig. 13.6 results in discontinuities. This is further displayed in Fig. 13.7. These discontinuities can be removed by designing the system such the transition between frequencies occurs with *continuous phase*, as shown in the left side of Fig. 13.6.

If the continuous phase system is designed such that the mark and space frequencies are synchronized with the input data frequency, then a partial orthogonality condition can be established. If two signals are orthogonal they can overlap in frequency and still be demodulated successfully. Partial orthogonality is established by setting the frequencies of the mark and space to odd multiples of ½ the input bit rate, i.e.,

$$f_s,\ f_m = \frac{n}{2t_b} = \frac{nf_b}{2}, \tag{11}$$

where the bit rate $f_b = 1/t_b$. Of course the n chosen for each frequency must be different from the other one.

When this condition is met, the partial orthogonality allows the spectra of the two frequencies to overlap. It is found that the spacing between the two frequencies can be reduced to

$$f_m - f_s \geq \frac{1}{2t_b}. \qquad (MSK) \tag{12}$$

Compare this with the minimum separation frequency given by Eq. 10 for the non-MSK system.

Since the carrier frequency is midway between the mark and space frequencies, we can use Eq. 12 to determine

$$f_m = f_c + \frac{1}{4t_b} \qquad and$$

$$f_s = f_c - \frac{1}{4t_b}.$$

The bandwidth extends from the first null below f_s (at $f_s - 1/t_b$) to the first null above f_m (at $f_m + 1/t_b$). The total bandwidth will then be

$$BW_{MSK} = \frac{1}{t_b} + \frac{1}{t_b} + \frac{1}{2t_b} = \frac{5}{2t_b}. \qquad \textbf{(14)}$$

Compare this with $4/t_b$ for standard FSK.

4. FSK Demodulation

Unlike FM, FSK can be demodulated using either coherent or non-coherent methods. Since there are only two transmitted frequencies, they can be separated using bandpass filters. Once separated the two signals can be envelope detected to recover the input data streams for both the marks and spaces. These can then recombined to re-form the original data sequence. This non-coherent demodulator is shown in Fig. 13.9.

A coherent demodulator is shown in Fig. 13.10 which actually contains two coherent demodulators, one for each of the two frequencies. Each operates as a standard coherent demodulator consisting of a multiplied coherent carrier followed by low pass filters.

The most common method of FSK demodulation is with a phase-locked loop. Just as with FM demodulation, the PLL delivers a dc error voltage to the output indicative of original input voltage.

C. PHASE SHIFT KEYING

Analog angle modulation consists of both frequency and phase modulations. In general they are indistinguishable with an analog input modulation signal. In digital systems, the input voltages are discrete, so the phase and frequency modulated signals do not appear identical. The digital phase modulated signal is called phase shift keying (PSK).

PSK is the most prevalent form of digital modulation, primarily because it has the best performance in the presence of noise. With better noise performance, the number of discrete representations can be increased from that of binary (2), to 4, 8, 16, or higher.

Like FSK, the amplitude of the PSK carrier is not modulated. But unlike FSK, the frequency of the carrier is also unchanged. Only the phase of the carrier is modified. Since the modulating signal is digital, the phases that the carrier can assume are discrete. The number of different phases that the carrier can assume is determined by the number of input states. We'll begin with two input states, i.e., a binary signal.

1. Binary PSK

Binary PSK (BPSK) is a popular type of PSK because it is easy to generate, re-generate, and it has the highest immunity to noise. To create BPSK we begin with the carrier signal given by Eq. 1. One state of the BPSK signal can be represented by the unmodulated carrier, and the other state can be its negative (180 degree phase change).

We again let the input signal consist of +1 volt to represent a mark and a –1 to represent a space, i.e., $v_m = \pm 1$. It is seen that BPSK can be created by the product of the input signal and the carrier, i.e.,

$$v_{BPSK}(t) = v_m(t) \cdot \cos(\omega_c t). \quad (15)$$

Notice that this is identical in form to that of DSB suppressed carrier (which gives us a hint as to its bandwidth requirements). The input signal and the resulting PSK modulated signal are shown in Fig. 13.11 in the text.

To determine the bandwidth required for passage of the BPSK signal, proceed as we did with FSK. First divide $v_{BPSK}(t)$ into its two components: the string of marks and the string of spaces. Taking first the marks, we see that its frequency spectrum is the same sin(f)/f which was pictured before for FSK. This spectrum is shifted out to f_c because it is multiplied by $\cos(\omega_c t)$. Again assuming a null-to-null bandwidth, the bandwidth extends from $f_c - 1/t_b$ to $f_c + 1/t_b$, or $BW_{Mark} = 2/t_b$.

Repeating the above procedure for the space inputs, we find the identical bandwidth again centered on the carrier frequency. The spectrum of the space inputs overlays that of the marks. Therefore, the bandwidth is seen to be $BW_{BPSK} = 2/t_b$.

Compare the bandwidth required for the BPSK system of that required for the FSK. With less bandwidth required, less noise is allowed to pass into the receiver, increasing the signal-to-noise ratio.

Because the bandwidth of the two signals overlap in frequency, non-coherent demodulation is not possible. The BPSK signal is demodulated using coherent demodulation.

2. M-ary PSK

We have examined modulating signals which contain two states, i.e., binary. The general term for the number of modulating states is M-ary where M is the number of states. For M-ary PSK the M indicates the number of discrete phases the modulated signal will be allowed to have.

Digital signals are first created in binary. This is because electronic circuits are easily configured to recognize the two states of a circuit–either on or off. Equating one state with a logical one and the other with a logical zero, a binary signal is created.

This binary signal can be used to modulate a carrier directly, resulting in binary FSK or BPSK, for example, or binary bits can be grouped together to form symbols. These new symbol-encoded signals can then be used to modulate a carrier, resulting in M-ary encoding.

With binary encoding, the output changes state each time the input changes state. With M-ary encoding higher than M = 2, the output changes state less often than that of the input, thereby increasing the effective value of t_b. If N is the number of bits that are input before the output state changes, then

$$N = \log_2 M. \tag{16}$$

This reduction in the number of output state changes has the desirable effect of reducing the required bandwidth of the modulated signal. The rate at which the output changes state is called the *baud* (not baud rate), which is in general different than the bps. For binary system the bps and the baud are the same.

3. Quarternary PSK

Another popular form of PSK is quarternary or quadrature PSK (QPSK). QPSK is M-ary PSK where M = 4. Putting M = 4 into Eq. 16 we see that N = 2, meaning that two input bits will be inputted into the modulator before the output is allowed to change state. The two-bit input pair required for state change is called a *dibit*. The rate of change of the output is reduced to half of the bit rate of the input, reducing the required system bandwidth by half.

With BPSK there were two output states, so that the output phase states were separated by 180 degrees. For QPSK, there are four states, and the allowed phases of the output signal are spaced 90 degrees apart.

4. M-PSK

PSK can be generated which allows phase states higher than four, as with QPSK. There are 8-PSK, 16-PSK, 32-PSK, etc. The number of allowed states, the M, can be inserted into Eq. 16 to determine the number of bits will be input into the modulator before the output is allowed to change state. For example, for 8-PSK, three bits (called a tribit) will be input before the output changes state. This reduces this system bandwidth to one-third that of the bit rate. These principles apply to higher PSK modes, as well.

5. Quadrature Amplitude Modulation

With normal PSK the amplitude of the carrier is kept constant, only the phase is changed. A special form of PSK also changes the amplitude in addition to the phase, called quadrature amplitude modulation (QAM).

The bandwidth requirements for an M-level QAM signal are the same as for the same level PSK signal. The advantage of QAM is it has higher immunity to noise.

D. BANDWIDTH EFFICIENCY

The bandwidth efficiency is a measure of how densely the information is contained in the transmitted bandwidth. To liken this concept to analog systems, recall that DSB AM and SSB AM carry the same amount of information. However, the DSB system requires twice the bandwidth to carry that information. The SSB system therefore has a higher bandwidth efficiency.

The measure of information capacity is the bit rate, in bits per second. The mathematical expression for the bandwidth efficiency is then

$$BW\ efficiency = \frac{f_b}{B}, \tag{17}$$

where B is the transmission bandwidth.

E. PROBABILITY OF ERROR

Errors in reception occur as a result of signal corruption due to noise or interference. In analog systems, the measure of receiver noise performance is the ratio of the average signal power to the noise power, SNR. This measure is often expressed in decibels. For a given system, a certain SNR is required to successfully demodulate the incoming analog signal.

In digital systems, a much tighter measure of noise performance is required. Not only is it required to have at least the minimum SNR for demodulation, but it is also necessary

to quantify the number of bit errors that will occur in a received signal. A bit error is the process of receiving a one when a zero was sent or vice versa. The cause of bit errors is again signal corruption due to noise.

Digital communications are very precise, delivering large amounts of data, free of errors. When transmitting data, a single bit error can cause problems whose severity ranges from inconvenient to calamitous. We therefore quantify receiver performance in the presence of noise for digital systems as the probability of a bit being in error, rather than with SNR.

The aim of a communications system designer is to achieve a zero probability of bit error. Since a probability of zero is difficult to attain, a very small number is adopted as the acceptable level of probability. This probability of error is a mathematical prediction and is given the symbol P(e) or sometimes P_e. Most systems strive for a value of P(e) less than 10^{-5}, i.e., a probability of one bit error in 10^5 bits transmitted.

Another term that is often seen to describe system bit errors is the bit error rate (BER). The terms BER and P(e) are often used interchangeably, but BER is actually a measured error quantity rather than predicted, as with P(e).

To compute P(e) we begin with the SNR. When computing the average signal and noise power we must agree on the signal and noise power sources. For the signal power in digital systems we normally refer to the power in the carrier and its sidebands, given the symbol C. For frequencies above HF, the relevant noise power is thermally generated. We will look at the carrier and noise powers.

To compute average power, divide the energy contained in a certain time interval of the signal divided by the length of the interval. If we choose the interval to be the length of a bit, t_b, then the average power is

$$C = \frac{E_b}{t_b} = E_b \cdot f_b \quad watts, \tag{18}$$

where E_b is the energy contained in the bit and f_b is the bit rate.

Assuming we are operating above HF, the average noise power is thermal, given by the well-known relationship

$$N = kTB \quad watts, \tag{19}$$

where N is the thermal noise power, k is Boltzman's constant (1.38×10^{-23} J/°K), T is the absolute temperature in degrees Kelvin (K), and B is the bandwidth in hertz. Thermal noise being considered white, the power in the noise is distributed evenly across the bandwidth. We can therefore compute the noise power density, N_0 (power per unit frequency), by dividing the average noise power by the bandwidth, i.e.,

$$N_0 = \frac{kTB}{B} = kT \quad \frac{watts}{hertz}. \tag{20}$$

It should be noted that units of watts/hertz are joules, units of energy.

Using Equations 18 and 19 the signal to noise ratio is seen to be

$$\frac{C}{N} = \frac{E_b \cdot f_b}{kTB} = \frac{E_b}{kT}\,\frac{f_b}{B} = \frac{E_b}{N_0}\,\frac{f_b}{B}, \tag{21}$$

where f_b/B is the bandwidth efficiency.

Since it is probability of bit error that we seek, we can isolate the term which concerns the ratio of the energy in the bit to the noise power density, giving

$$\frac{E_b}{N_0} = \frac{C}{N}\,\frac{B}{f_b}. \tag{22}$$

The terms on the right side of the equation are normally known quantities, so that E_b/N_0 is readily computed. Notice that if the bandwidth is equal to the bit rate, E_b/N_0 is equal to C/N. This quantity is often converted to decibels,

$$\left(\frac{E_b}{N_0}\right)_{dB} = 10\,\log\left(\frac{C}{N}\right) + 10\,\log\left(\frac{B}{f_b}\right). \tag{23}$$

The ratio of E_b to N_0 is an energy ratio which represents the amount of bit signal energy with respect to the noise energy. Whether the bit will be received correctly is directly proportional to this ratio. Conversely, the probability that the bit will be received in error, i.e., P(e), is inversely proportional to this ratio. The exact relationship of P(e) to E_b/N_0 varies, depending upon the type of modulation, as given in the Figures 13.26 through 13-29.